*STATISTICS IN TRANSITION new series, June 2019*
*Vol. 20, No. 2, pp. 123–138, DOI 10.21307/stattrans-2019-018*

# VARIABLE SELECTION IN MULTIVARIATE FUNCTIONAL DATA CLASSIFICATION

**Tomasz Górecki**[1], **Mirosław Krzyśko**[2], **Waldemar Wołyński**[3]

## ABSTRACT

A new variable selection method is considered in the setting of classification with multivariate functional data (Ramsay and Silverman (2005)). The variable selection is a dimensionality reduction method which leads to replace the whole vector process, with a low-dimensional vector still giving a comparable classification error. Various classifiers appropriate for functional data are used. The proposed variable selection method is based on functional distance covariance (dCov) given by Székely and Rizzo (2009, 2012) and the Hilbert-Schmidt Independent Criterion (HSIC) given by Gretton et al. (2005). This method is a modification of the procedure given by Kong et al. (2015). The proposed methodology is illustrated with a real data example.

**Key words:** multivariate functional data, variable selection, dCov, HSIC, classification.

## 1. Introduction

In recent years, much attention has been paid to methods for representing data as functions or curves. Such data are known in the literature as functional data (Ramsay and Silverman (2005), Horváth and Kokoszka (2012)). Applications of functional data can be found in various fields, including medicine, economics, meteorology and many others. In many applications there is a need to use statistical methods for objects characterized by multiple variables observed at many time points (doubly multivariate data). Such data are called multivariate functional data. In this paper we focus on the classification problem for multivariate functional data. In many cases, in the classification procedures, the number of predictors $p$ is significantly greater than the sample size $n$. Thus, it is natural to assume that only a small number of predictors are relevant to response $Y$.

Various basic classification methods have also been adapted to functional data, such as linear discriminant analysis (Hastie et al. (1995)), logistic regression (Rossi

[1]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. E-mail: tomasz.gorecki@amu.edu.pl. ORCID ID: https://orcid.org/0000-0002-9969-5257.

[2]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. Interfaculty Institute of Mathematics and Statistics, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Poland. E-mail: mkrzysko@amu.edu.pl. ORCID ID: https://orcid.org/0000-0001-0075-4432.

[3]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland.
E-mail: wolynski@amu.edu.pl. ORCID ID: https://orcid.org/0000-0002-0777-9163.

et al. (2002)), penalized optimal scoring (Ando (2009)), $k$nn (Ferraty and Vieu (2003)), SVM (Rossi and Villa (2006)), and neural networks (Rossi et al. (2005)). Moreover, the combining of classifiers has been extended to functional data (Ferraty and Vieu (2009)). Górecki et al. (2016) adapted multivariate regression models to the classification of multivariate functional data. Gretton et al. (2005) defined the measure of dependence between random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ called the Hilbert-Schmidt Independence Criterion (HSIC) and proved that this measure is equal to zero if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent to each other when using universal kernels, such as the Gaussian kernels. Based on the idea of HSIC between two random vectors, we introduced the HSIC between two random processes.

Székely et al. (2007), Székely and Rizzo (2009, 2012, 2013) defined the measures of dependence between random vectors: the distance covariance (dCov). These authors showed that for all random variables with finite first moments, dCov generalizes the idea of covariance in two ways. Firstly, this coefficient can be applied when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are of any dimensions and not only for the simple case where $p = q = 1$. Secondly, dCov is equal to zero if and only if there is independence between the random vectors. Indeed, the distance covariance measures a linear relationship and can be equal to 0 even when the variables are related. Based on the idea of the distance covariance between two random vectors, we introduced the functional distance covariance between two random processes. We select a set of important predictors with a large value of functional distance covariance or functional Hilbert-Schmidt Independent Criterion. Our selection procedure is a modification of the procedure given by Kong et al. (2015).

An entirely different approach to the variable selection in functional data classification is presented by Berrendero et al. (2016). It is clear that variable selection has, at least, an advantage when compared with other dimension reduction methods (functional principal component analysis (FPCA), see Górecki et al. (2014), Jacques and Preda (2014), functional partial least squares (FPLS) methodology, see Delaigle and Hall (2012), and other methods) based on general projections: the output of any variable selection method is always directly interpretable in terms of the original variables, provided that the required number $d$ of the selected variables is not too large.

The rest of this paper is organized as follows. In Section 2 we present the classification procedures used through the paper. In Section 3 we present the problem of representing functional data by orthonormal basis functions. In Section 4, we define a functional distance covariance. In Section 5 we define a functional HSIC. In Section 6 we propose a variable selection procedure based on the functional distance covariance and on HSIC. In Section 7 we illustrate the proposed methodology through a real data example. We conclude in Section 8.

## 2. Classifiers

The classification problem involves determining a procedure by which a given object can be assigned to one of $q$ populations based on observation of $p$ features of that

object.

The object being classified can be described by a random pair $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^\top \in \mathbb{R}^p$ and $Y \in \{1, \ldots, q\}$. An automated classifier can be viewed as a method of estimating the posterior probability of membership in groups. For a given $\boldsymbol{X}$, a reasonable strategy is to assign $\boldsymbol{X}$ to that class with the highest posterior probability. This strategy is called the Bayes' rule classifier.

### 2.1. Linear and quadratic discriminant classifiers

Now we make the Bayes' rule classifier more specific by the assumption that all multivariate probability densities are multivariate normal having arbitrary mean vectors and a common covariance matrix. We shall call this model the linear discriminant classifier (LDC). Assuming that class-covariance matrices are different, we obtain quadratic discriminant classifier (QDC).

### 2.2. Naive Bayes classifier

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with independence assumptions. When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a one-dimensional normal distribution or we estimate density by kernel method.

### 2.3. $k$-nearest neighbour classifier

Most often we do not have sufficient knowledge of the underlying distributions. One of the important nonparametric classifiers is a $k$-nearest neighbour classifier ($k$NN classifier). Objects are assigned to the class having the majority in the $k$ nearest neighbours in the training set.

### 2.4. Multinomial logistic regression

It is a classification method that generalizes logistic regression to multiclass problem using one vs. all approach.

## 3. Functional data

We now assume that the object being classified is described by a $p$-dimensional random process $\boldsymbol{X} = (X_1, X_2, ..., X_p)^\top \in L_2^p(I)$, where $L_2(I)$ is the Hilbert space of square-integrable functions, and $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{0}$.

Moreover, assume that the $k$th component of the vector $\boldsymbol{X}$ can be represented by a finite number of orthonormal basis functions $\{\varphi_b\}$

$$X_k(t) = \sum_{b=0}^{B_k} \alpha_{kb}\varphi_b(t),\ t \in I,\ k = 1, \ldots, p,$$

where $\alpha_{k0}, \alpha_{k1}, \ldots, \alpha_{kB_k}$ are the unknown coefficients.

Let $\boldsymbol{\alpha} = (\alpha_{10}, \ldots, \alpha_{1B_1}, \ldots, \alpha_{p0}, \ldots, \alpha_{pB_p})^\top \in \mathbb{R}^{K+p}$, $K = B_1 + \cdots + B_p$ and

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\varphi}_1^\top(t) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varphi}_2^\top(t) & \ldots & \mathbf{0} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\varphi}_p^\top(t) \end{bmatrix},$$

where $\boldsymbol{\varphi}_k(t) = (\varphi_0(t), ..., \varphi_{B_k}(t))^\top$, $k = 1, ..., p$.

Using the above matrix notation, the process $\boldsymbol{X}$ can be represented as:

$$\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\alpha}, \tag{1}$$

where $\mathrm{E}(\boldsymbol{\alpha}) = \mathbf{0}$. This means that the realizations of the process $\boldsymbol{X}$ are in finite dimensional subspace of $L_2^p(I)$. We will denote this subspace by $\mathscr{L}_2^p(I)$.

We can estimate the vector $\boldsymbol{\alpha}$ on the basis of $n$ independent realizations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ of the random process $\boldsymbol{X}$ (functional data). We will denote this estimator by $\hat{\boldsymbol{\alpha}}$.

Typically data are recorded at discrete moments in time. Let $x_{kj}$ denote an observed value of the feature $X_k$, $k = 1, 2, \ldots, p$ at the $j$th time point $t_j$, where $j = 1, 2, ..., J$. Then our data consist of the $pJ$ pairs $(t_j, x_{kj})$. These discrete data can be smoothed by continuous functions $x_k$ and $I$ is a compact set such that $t_j \in I$, for $j = 1, ..., J$.

Details of the process of transformation of discrete data to functional data can be found in Ramsay and Silverman (2005) or in Górecki et al. (2014).

## 4. Distance covariance (dCov)

For the jointly distributed random process $\boldsymbol{X} \in L_2^p(I)$ and the random vector $\boldsymbol{Y} \in \mathbb{R}^q$, let

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l}, \boldsymbol{m}) = \mathrm{E}\{\exp[i\langle \boldsymbol{l}, \boldsymbol{X}\rangle + i\langle \boldsymbol{m}, \boldsymbol{Y}\rangle_q]\}$$

be the joint characteristic function of $(\boldsymbol{X}, \boldsymbol{Y})$, where

$$\langle \boldsymbol{l}, \boldsymbol{X}\rangle = \int_I \boldsymbol{l}'(t)\boldsymbol{X}(t)dt$$

and

$$\langle \boldsymbol{m}, \boldsymbol{Y}\rangle = \boldsymbol{m}'\boldsymbol{Y}.$$

Moreover, we define the marginal characteristic functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as follows: $f_{\boldsymbol{X}}(\boldsymbol{l}) = f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l}, \mathbf{0})$ and $f_{\boldsymbol{Y}}(\boldsymbol{m}) = f_{\boldsymbol{X},\boldsymbol{Y}}(\mathbf{0}, \boldsymbol{m})$.

Here, for generality, we assume that $\boldsymbol{Y} \in \mathbb{R}^q$, although the label $Y$ in the classification problem is a random variable, with values in $\{1, \ldots, q\}$. Label $\boldsymbol{Y}$ has to be transformed into the label vector $\boldsymbol{Y} = (Y_1, \ldots, Y_q)'$, where $Y_i = 1$ for $i = 1, \ldots, q$ if $\boldsymbol{X}$ belongs to class $i$, and 0 otherwise.

Now, let us assume that $\boldsymbol{X} \in \mathscr{L}_2^p(I)$. Then, the process $\boldsymbol{X}$ has the representation (1).

In this case, we may assume (Ramsay and Silverman (2005)) that the vector weight function $\boldsymbol{l}$ and the process $\boldsymbol{X}$ are in the same space, i.e. the function $\boldsymbol{l}$ can be written in the form

$$\boldsymbol{l}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\lambda}, \tag{2}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{K+p}$.

Hence

$$\langle \boldsymbol{l}, \boldsymbol{X} \rangle = \int_I \boldsymbol{l}'(t)\boldsymbol{X}(t)dt = \boldsymbol{\lambda}'[\int_I \boldsymbol{\Phi}'(t)\boldsymbol{\Phi}(t)dt]\boldsymbol{\alpha} = \boldsymbol{\lambda}'\boldsymbol{\alpha},$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ are vectors occurring in the representations (1) and (2) of the process $\boldsymbol{X}$ and function $\boldsymbol{l}$, and

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l},\boldsymbol{m}) = \mathrm{E}\{\exp[i\boldsymbol{\lambda}'\boldsymbol{\alpha} + i\boldsymbol{m}'\boldsymbol{Y}]\} = f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m}),$$

where $f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m})$ is the joint characteristic function of the pair of random vectors $(\boldsymbol{\alpha},\boldsymbol{Y})$.

On the basis of the idea of distance covariance between two random vectors (Székely et al. (2007)), we can introduce functional distance covariance between random process $\boldsymbol{X}$ and random vector $\boldsymbol{Y}$.

**Definition 1.** *A nonnegative number* $\mathrm{dCov}(\boldsymbol{X},\boldsymbol{Y})$ *defined by*

$$\mathrm{dCov}(\boldsymbol{X},\boldsymbol{Y}) = \mathrm{dCov}(\boldsymbol{\alpha},\boldsymbol{Y}),$$

*where*

$$\mathrm{dCov}^2(\boldsymbol{\alpha},\boldsymbol{Y}) = \frac{1}{C_{K+p}C_q}\int_{\mathbb{R}^{K+p+q}} \frac{|f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m}) - f_{\boldsymbol{\alpha}}(\boldsymbol{\lambda})f_{\boldsymbol{Y}}(\boldsymbol{m})|^2}{\|\boldsymbol{\lambda}\|_{K+p}^{K+p+1}\|\boldsymbol{m}\|_q^{q+1}} d\boldsymbol{\lambda}\, d\boldsymbol{m},$$

*and* $|z|$ *denotes the modulus of* $z \in \mathbb{C}$, $\|\boldsymbol{\lambda}\|_{K+p}$, $\|\boldsymbol{m}\|_q$ *the standard Euclidean norms on the corresponding spaces* $V$ *chosen to produce scale free and rotation invariant measure that does not go to zero for dependent random vectors, and*

$$C_r = \frac{\pi^{\frac{1}{2}(r+1)}}{\Gamma(\frac{1}{2}(r+1))}$$

*is half the surface area of the unit sphere in* $\mathbb{R}^{r+1}$*, is called a functional distance covariance between the random process* $\boldsymbol{X}$ *and the random vector* $\boldsymbol{Y}$*.*

We can estimate functional distance covariance using data set $\boldsymbol{S} = \{(\hat{\boldsymbol{\alpha}}_1,\boldsymbol{y}_1),\ldots,(\hat{\boldsymbol{\alpha}}_n,\boldsymbol{y}_n)\}$.

Let

$$\bar{\boldsymbol{\alpha}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\alpha}}_k, \quad \bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_k,$$
$$\tilde{\boldsymbol{\alpha}}_k = \hat{\boldsymbol{\alpha}}_k - \bar{\boldsymbol{\alpha}}, \quad \tilde{\boldsymbol{y}}_k = \boldsymbol{y}_k - \bar{\boldsymbol{y}},\ k = 1,\ldots,n$$

and

$$\boldsymbol{A} = (a_{kl}), \quad \boldsymbol{B} = (b_{kl}),$$
$$\tilde{\boldsymbol{A}} = (A_{kl}), \quad \tilde{\boldsymbol{B}} = (B_{kl}),$$

where

$$a_{kl} = \|\hat{\boldsymbol{\alpha}}_k - \hat{\boldsymbol{\alpha}}_l\|_{K+p}, \quad b_{kl} = \|\boldsymbol{y}_k - \boldsymbol{y}_l\|_q,$$
$$A_{kl} = \|\tilde{\boldsymbol{\alpha}}_k - \tilde{\boldsymbol{\alpha}}_l\|_{K+p}, \quad B_{kl} = \|\tilde{\boldsymbol{y}}_k - \tilde{\boldsymbol{y}}_l\|_q, \quad k,l = 1,\ldots,n.$$

Hence

$$\tilde{\boldsymbol{A}} = \boldsymbol{HAH}, \quad \tilde{\boldsymbol{B}} = \boldsymbol{HBH},$$

where

$$\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$$

is the centring matrix.

On the basis of the result of Székely et al. (2007), we have

$$\mathrm{dCov}(\boldsymbol{S}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}.$$

## 5. Hilbert-Schmidt Independent Criterion (HSIC)

Let $\boldsymbol{\phi}$ be a mapping from $L_2^p$ to an inner product feature space $\mathscr{H}$, and $\boldsymbol{\psi}$ be a mapping from $R^q$ to an inner product feature space $\mathscr{G}$.

**Definition 2.** *The cross-covariance operator $\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}}: \mathscr{G} \to \mathscr{H}$ is a linear operator defined as*

$$\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}} = \mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}[\boldsymbol{\phi}(\boldsymbol{X}) \otimes \boldsymbol{\psi}(\boldsymbol{Y})] - \mu_{\boldsymbol{X}} \otimes \mu_{\boldsymbol{Y}},$$

*for all $f \in \mathscr{H}$ and $g \in \mathscr{G}$, where the tensor product operator $f \otimes g: \mathscr{G} \to \mathscr{H}$, $f \in \mathscr{H}$, $g \in \mathscr{G}$, is defined as*

$$(f \otimes g)h = f\langle g,h\rangle_{\mathscr{G}}, \text{ for all } h \in \mathscr{G}.$$

This is a generalization of the cross-covariance matrix between random vectors.

Moreover, by the definition of the Hilbert-Schmidt (HS) norm, we can compute the HS norm of $f \otimes g$ via

$$\|f \otimes g\|_{HS}^2 = \|f\|_{\mathscr{H}}^2 \|g\|_{\mathscr{G}}^2.$$

Gretton et al. (2005) defined the Hilbert-Schmidt Independence Criterion (HSIC) in the following way:

**Definition 3.** *Hilbert-Schmidt Independence Criterion (HSIC) is the squared Hilbert-Schmidt norm of the cross-covariance operator*

$$\mathrm{HSIC}(\boldsymbol{X},\boldsymbol{Y}) = \|\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}}\|_{HS}^2.$$

Now, let

$$k\colon \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$$

be a kernel function on $\mathbb{R}^p$.

This raises an interesting question: given a function of two variables $k(\boldsymbol{x},\boldsymbol{x}')$, does there exist a function $\boldsymbol{\phi}$ such that $k(\boldsymbol{x},\boldsymbol{x}') = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}') \rangle_{\mathscr{H}}$? The answer is provided by Mercer's theorem (1909), which says, roughly, that if $k$ is positive semi-definite then such a $\boldsymbol{\phi}$ exists.

Often, we will not know $\boldsymbol{\phi}$, but a kernel function $k$, which encodes the inner product in $\mathscr{H}$, instead.

Popular positive semi-definite kernel functions on $\mathbb{R}^p$ include the polynomial kernel of degree $d > 0$, $k(\boldsymbol{x},\boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^d$, the Gaussian kernel $k(\boldsymbol{x},\boldsymbol{x}') = \exp(-\lambda \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$, $\lambda > 0$, and the Laplace kernel $k(\boldsymbol{x},\boldsymbol{x}') = \exp(-\lambda \|\boldsymbol{x} - \boldsymbol{x}'\|)$, $\lambda > 0$. In this paper we use, the Gaussian kernel.

A feature mapping $\boldsymbol{\phi}$ is centred by subtracting from it its expectation, that is transforming $\boldsymbol{\phi}(\boldsymbol{x})$ to $\tilde{\boldsymbol{\phi}}(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}) - \mathrm{E}_{\boldsymbol{X}}[\boldsymbol{\phi}(\boldsymbol{X})]$. Centring a positive semi-definite kernel function $k$ consists in centring in the feature mapping $\boldsymbol{\phi}$ associated to $k$. Thus, the centred kernel $\tilde{k}$ associated to $k$ is defined by

$$\begin{aligned}\tilde{k}(\boldsymbol{x},\boldsymbol{x}') &= \langle \boldsymbol{\phi}(\boldsymbol{x}) - \mathrm{E}_{\boldsymbol{X}}[\boldsymbol{\phi}(\boldsymbol{X})], \boldsymbol{\phi}(\boldsymbol{x}') - \mathrm{E}_{\boldsymbol{X}'}[\boldsymbol{\phi}(\boldsymbol{X}')] \rangle \\ &= k(\boldsymbol{x},\boldsymbol{x}') - \mathrm{E}_{\boldsymbol{X}}[k(\boldsymbol{X},\boldsymbol{x}')] - \mathrm{E}_{\boldsymbol{X}'}[k(\boldsymbol{x},\boldsymbol{X}')] + \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[k(\boldsymbol{X},\boldsymbol{X}')],\end{aligned}$$

assuming the expectations exist. Here, the expectation is taken over independent copies $\boldsymbol{X}, \boldsymbol{X}'$. We see that, $\tilde{k}$ is also a positive semi-definite kernel. Note also that for a centred kernel $\tilde{k}$, $\mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[\tilde{k}(\boldsymbol{X},\boldsymbol{X}')] = 0$, that is, centring the feature mapping implies centring the kernel function.

Let $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ be a finite subset of $\mathbb{R}^p$. A feature mapping $\boldsymbol{\phi}$ is centred by subtracting from it its empirical expectation, i.e. leading to $\tilde{\boldsymbol{\phi}}(\boldsymbol{x}_i) = \boldsymbol{\phi}(\boldsymbol{x}_i) - \overline{\boldsymbol{\phi}}$, where $\overline{\boldsymbol{\phi}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}(\boldsymbol{x}_i)$. The kernel matrix $\boldsymbol{K} = (K_{ij})$ associated to the kernel function $k$ and the set $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ is centred by replacing it with $\widetilde{\boldsymbol{K}} = (\widetilde{K}_{ij})$ defined for all $i, j =$

$1,2,\ldots,n$ by

$$\tilde{K}_{ij} = K_{ij} - \frac{1}{n}\sum_{i=1}^{n} K_{ij} - \frac{1}{n}\sum_{j=1}^{n} K_{ij} + \frac{1}{n^2}\sum_{i,j=1}^{n} K_{ij},$$

where $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1, \ldots, n$.

The centred kernel matrix $\widetilde{\boldsymbol{K}}$ is a positive semi-definite matrix. Also, as with the kernel function $\frac{1}{n^2}\sum_{i,j}^{n} \tilde{K}_{ij} = 0$.

Let $\langle \cdot, \cdot \rangle_F$ denote the Frobenius product and $\| \cdot \|_F$ the Frobenius norm defined for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}$ by

$$\begin{aligned} \langle \boldsymbol{A}, \boldsymbol{B} \rangle_F &= \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B}), \\ \|\boldsymbol{A}\|_F &= (\langle \boldsymbol{A}, \boldsymbol{A} \rangle_F)^{1/2}. \end{aligned}$$

Then, for any kernel matrix $\boldsymbol{K} \in \mathbb{R}^{n \times n}$, the centred kernel matrix $\widetilde{\boldsymbol{K}}$ can be expressed as follows (Schölkopf et al.(1998)):

$$\widetilde{\boldsymbol{K}} = \boldsymbol{H}\boldsymbol{K}\boldsymbol{H},$$

where $\boldsymbol{H}$ ia a centering matrix.

Since $\boldsymbol{H}$ is the idempotent matrix ($\boldsymbol{H}^2 = \boldsymbol{H}$), then we get for any two kernel matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ based on the subset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of $\mathbb{R}^p$ and the subset $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of $\mathbb{R}^q$, respectively,

$$\langle \widetilde{\boldsymbol{K}}, \widetilde{\boldsymbol{L}} \rangle_F = \langle \boldsymbol{K}, \widetilde{\boldsymbol{L}} \rangle_F = \langle \widetilde{\boldsymbol{K}}, \boldsymbol{L} \rangle_F.$$

We may express HSIC in terms of kernel functions (Gretton et al. (2005)):

$$\begin{aligned} \operatorname{HSIC}(\boldsymbol{X}, \boldsymbol{Y}) = {} & \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}',\boldsymbol{Y},\boldsymbol{Y}'}[k(\boldsymbol{X}, \boldsymbol{X}')l(\boldsymbol{Y}, \boldsymbol{Y}')] \\ & + \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[k(\boldsymbol{X}, \boldsymbol{X}')]\,\mathrm{E}_{\boldsymbol{Y},\boldsymbol{Y}'}[l(\boldsymbol{Y}, \boldsymbol{Y}')] \\ & - 2\,\mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}[\mathrm{E}_{\boldsymbol{X}'}[k(\boldsymbol{X}, \boldsymbol{X}')]\,\mathrm{E}_{\boldsymbol{Y}'}[l(\boldsymbol{Y}, \boldsymbol{Y}')]]. \end{aligned}$$

Here, $\mathrm{E}_{\boldsymbol{X},\boldsymbol{X}',\boldsymbol{Y},\boldsymbol{Y}'}$ denotes the expectation over independent pairs $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}', \boldsymbol{Y}')$.

Let

$$k^\star : \mathscr{L}_2^p(I) \times \mathscr{L}_2^p(I) \to \mathbb{R}$$

be a kernel function on $\mathscr{L}_2^p(I)$. For the multivariate functional data the Gaussian kernel has the form:

$$k^\star(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\lambda \|\boldsymbol{x} - \boldsymbol{x}'\|^2),\ \lambda > 0.$$

From the orthonormality of the basis functions, we have:

$$\begin{aligned}\|\boldsymbol{x}-\boldsymbol{x}'\|^2 &= \int_I (\boldsymbol{x}(t)-\boldsymbol{x}'(t))^\top(\boldsymbol{x}(t)-\boldsymbol{x}'(t))dt \\ &= \|\hat{\boldsymbol{\alpha}}-\hat{\boldsymbol{\alpha}}'\|^2.\end{aligned}$$

Hence

$$k^\star(\boldsymbol{x},\boldsymbol{x}') = k(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\alpha}}'),$$

where $\hat{\boldsymbol{\alpha}}_1,\ldots,\hat{\boldsymbol{\alpha}}_n$ are vectors occurring in the representation (1).

**Definition 4.** *The empirical HSIC for functional data is defined as*

$$\mathrm{HSIC}(S^\star) = \frac{1}{n^2}\langle \boldsymbol{K}^\star,\boldsymbol{L}^\star\rangle_F,$$

*where* $S^\star = \{(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n)\}$, $\boldsymbol{K}^\star$ *and* $\boldsymbol{L}^\star$ *are kernel matrices based on the subsets* $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$, *and* $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n\}$ *of* $\mathscr{L}_2^p(I)$ *and* $\mathbb{R}^q$, *respectively.*

But $\boldsymbol{K}^\star = \boldsymbol{K}$, where $\boldsymbol{K}$ is the kernel matrix of size $n\times n$, which has its $(i,j)$th element $K_{ij}$ given by $K_{ij} = k(\hat{\boldsymbol{\alpha}}_i,\hat{\boldsymbol{\alpha}}_j)$. $\boldsymbol{L}$ is the kernel matrix of size $n\times n$, which has its $(i,j)$th element $L_{ij}$ given by $L_{ij} = l(\boldsymbol{y}_i,\boldsymbol{y}_j)$.

Hence

$$\mathrm{HSIC}(S^\star) = \mathrm{HSIC}(S),$$

where $S = \{(\hat{\boldsymbol{\alpha}}_1,\boldsymbol{y}_1),\ldots,(\hat{\boldsymbol{\alpha}}_n,\boldsymbol{y}_n)\}$.

## 6. Variable selection based on the fuctional dCov and the functional HSIC

In this Section we propose the selection procedure built on the functional dCov and the functional HSIC. Let $\boldsymbol{Y} = (Y_1,\ldots,Y_q)'$, be the response vector, and $\boldsymbol{X} = (X_1,\ldots,X_p)'$ be the predictor $p$-dimensional process. Assume that only a small number of predictors are relevant to $\boldsymbol{Y}$. We will define an irrelevant variable to be one whose value is statistically independent of label vector $\boldsymbol{Y}$ and of the other variables $X_1,\ldots,X_p$.

We select a set of important predictors with large functional $\mathrm{dCov}(\boldsymbol{S})$ or with large functional $\mathrm{HSIC}(\boldsymbol{S})$.

We utilize the functional dCov because it allows for arbitrary relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$, regardless of whether it is linear or nonlinear. We would like an assurance that irrelevant variables do not increase dCov. Kong et al. (2015) proved the following theorem.

**Theorem 1.** *Let* $\boldsymbol{Z} = (\boldsymbol{X}^\top, X_{p+1})^\top$, *where* $X_{p+1}$ *is an irrelevant variable. Then*

$$\mathrm{dCov}(\boldsymbol{Z},\boldsymbol{Y}) \le \mathrm{dCov}(\boldsymbol{X},\boldsymbol{Y}).$$

Gretton et al. (2005) proved that $\mathrm{HSIC}(\boldsymbol{X},\boldsymbol{Y}) = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent of each other. This is the direct motivation why we may also choose HSIC to measure the dependence. For the Gaussian kernel the following result is true.

**Theorem 2.** *Let* $\boldsymbol{Z} = (\boldsymbol{X}^\top, X_{p+1})^\top$*, where* $X_{p+1}$ *is an irrelevant variable. Then*

$$\mathrm{HSIC}(\boldsymbol{Z},\boldsymbol{Y}) \leq \mathrm{HSIC}(\boldsymbol{X},\boldsymbol{Y}).$$

*Proof.* Since the variable $X_{p+1}$ is independent of the label vector $\boldsymbol{Y}$ and the other variables $X_1,\ldots,X_p$, functions of these variables are also independent.

Hence

$$\begin{aligned}
\mathrm{HSIC}(\boldsymbol{Z},\boldsymbol{Y}) &= \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}',\boldsymbol{Y},\boldsymbol{Y}'}[k(\boldsymbol{Z},\boldsymbol{Z}')l(\boldsymbol{Y},\boldsymbol{Y}')] + \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[k(\boldsymbol{Z},\boldsymbol{Z}')]\,\mathrm{E}_{\boldsymbol{Y},\boldsymbol{Y}'}[l(\boldsymbol{Y},\boldsymbol{Y}')] \\
&\quad - 2\,\mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}\{\mathrm{E}_{\boldsymbol{X}'}[k(\boldsymbol{Z},\boldsymbol{Z}')]\,\mathrm{E}_{\boldsymbol{Y}'}[l(\boldsymbol{Y},\boldsymbol{Y}')]\} \\
&= \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}',\boldsymbol{Y},\boldsymbol{Y}'}[k(\boldsymbol{X},\boldsymbol{X}')\exp(-\lambda(X_{p+1}-X'_{p+1})^2)l(\boldsymbol{Y},\boldsymbol{Y}')] \\
&\quad + \mathrm{E}_{\boldsymbol{X},\boldsymbol{X}'}[k(\boldsymbol{X},\boldsymbol{X}')]\exp(-\lambda(X_{p+1}-X'_{p+1})^2)\,\mathrm{E}_{\boldsymbol{Y},\boldsymbol{Y}'}[l(\boldsymbol{Y},\boldsymbol{Y}')] \\
&\quad - 2\,\mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}\{\mathrm{E}_{\boldsymbol{X}'}[k(\boldsymbol{X},\boldsymbol{X}')\exp(-\lambda(X_{p+1}-X'_{p+1})^2)]\,\mathrm{E}_{\boldsymbol{Y}'}[l(\boldsymbol{Y},\boldsymbol{Y}')]\} \\
&= \mathrm{HSIC}(\boldsymbol{X},\boldsymbol{Y})\exp(-\lambda(X_{p+1}-X'_{p+1})^2) \leq \mathrm{HSIC}(\boldsymbol{X},\boldsymbol{Y}),
\end{aligned}$$

because $\exp(-\lambda(X_{p+1}-X'_{p+1})^2) \leq 1$, for $\lambda > 0$. □

The functional dCov and functional HSIC also permit univariate and multivariate response. Thus, this procedure is completely model-free.

We implemented the above theorems as a stopping rule in the selections of responses. The procedure took the following steps:

1. Calculate marginal functional dCov or functional HSIC for $X_k$, $k = 1,\ldots,p$ with the response $\boldsymbol{Y}$.
2. Rank the variables in decreasing order of the selected measure. Denote the ordered predictors as $X_{(1)},X_{(2)},\ldots,X_{(p)}$. Start with $\boldsymbol{X}_S = \{X_{(1)}\}$.
3. For $k$ from 2 to $p$, keep adding $X_{(k)}$ to $\boldsymbol{X}_S$ if $\mathrm{dCov}(\boldsymbol{X}_S,\boldsymbol{Y})$ or $\mathrm{HSIC}(\boldsymbol{X}_S,\boldsymbol{Y})$ does not decrease. Stop otherwise.

## 7. Example

The described method was employed here to select the variables (pillars) in the classification problem of 115 countries in the period 2008-2017. Table 1 describes the variables (pillars) used in the analysis.

**Table 1.** Variables (pillars) used in analysis, 2008-2017

| No. | Variable (pillar) |
|---|---|
| 1. | Institutions |
| 2. | Infrastructure |
| 3. | Macroeconomic environment |
| 4. | Health and primary education |
| 5. | Higher education and training |
| 6. | Goods market efficiency |
| 7. | Labour market efficiency |
| 8. | Financial market development |
| 9. | Technological readiness |
| 10. | Market size |
| 11. | Business sophistication |
| 12. | Innovation |

For this purpose, the use was made of data published by the World Economic Forum (WEF) in its annual reports (`http://www.weforum.org`). Those are comprehensive data, describing exhaustively various socio-economic conditions or spheres of individual states. WEF experts have divided discussed countries into five groups (Figure 1).

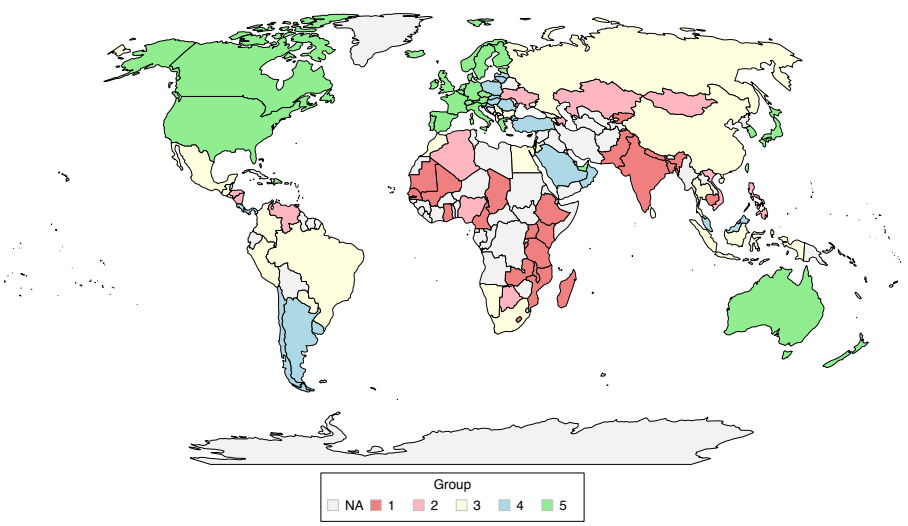


Figure 1: 115 countries used in the analysis

The data were transformed into functional data. Calculations were performed using the Fourier basis. In view of a small number of time periods, for each variable the maximum number of basis components was taken to be equal to five.

In the next step we applied the method of selecting variables described earlier (we stopped the procedure if the increase in the selected measure was less than 0.05). In such a way we obtained 5 variables (Figure 2 and Figure 3).

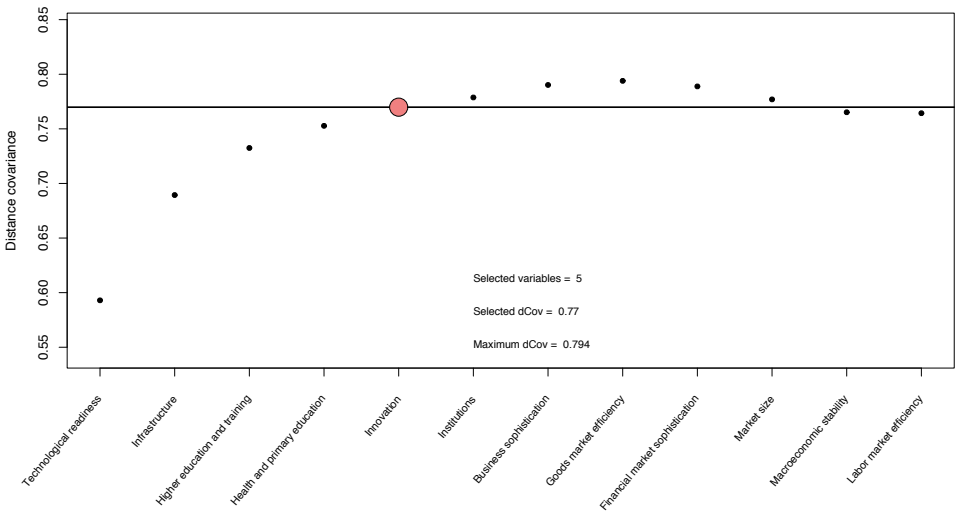


Figure 2: Variables selection for functional dCov

Next, we applied the described classifiers to reduced functional data and to full functional data. To estimate the error rate of the classifiers we used LOO CV (leave-one-out cross validation) method. The results are in Table 2.

**Table 2.** Classification accuracy (in %)

| Classifier | Selected variables (5) | All variables (12) |
|---|---|---|
| LDC | 71.30 | 66.09 |
| kNN ($k = 1, \ldots, 8$) | 77.39 | 71.30 |
| Naive Bayes (normal) | 69.57 | 65.22 |
| Naive Bayes (kernel) | 67.83 | 62.61 |
| Logistic regression | 60.87 | 56.52 |

We can observe that the error rate decreases if we reduce our data set. We can also notice that the order of classifiers stays unchanged (the best classifier for full data is kNN, and the same is the best for reduced data).

During the calculations we used `R` (R Core Team (2018)) software and `caret` (Kuhn (2018)), `energy` (Rizzo and Székely (2018)) and `fda` (Ramsay et al. (2018)) packages.

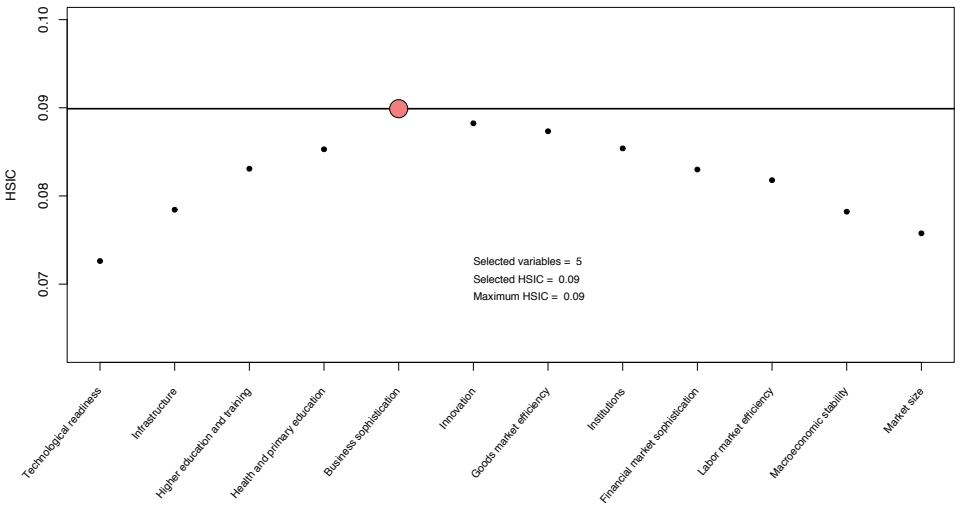


Figure 3: Variables selection for functional HSIC

## 8. Conclusions

The paper introduces variable selection for classification of multivariate functional data. The use of functional distance covariance or functional HSIC as a tool to reduce dimensionality of data set suggests that the technique provides useful results for classification of multivariate functional data. For analysed data set only five from twelve variables were included in the final model. We realize that the classification accuracy could drop slightly. However, we expect that this drop should be reasonable and in return we could gain a considerable amount of computation time.

In practice, it is important not to depend entirely on variable selection criteria because none of them works well under all conditions. So, our approach could be seen as a competitive to other variable selection methods and the full model without variables reduction. Finally, the researcher needs to evaluate the models using various diagnostic procedures.

## REFERENCES

ANDO, T., (2009). Penalized optimal scoring for the classification of multi-dimensional functional data, Statistical Methodology, 6, pp. 565–576.

BERRENDERO, J. R., CUEVAS, A., TORRECILLA, J. L., (2016). Variable selection in functional data classification: a maxima-hunting proposal, Statistica Sinica, 26 (2), pp. 619–638.

DELAIGLE, A., HAAL, P., (2012). Methodology and theory for partial least squares applied to functional data. Annals of Statistics, 40, pp. 322–352.

FERRATY, F., VIEU, P., (2003). Curve discrimination. A nonparametric functional approach. Computational Statistics & Data Analysis, 44, pp. 161–173.

FERRATY, F., VIEU, P., (2009). Additive prediction and boosting for functional data. Computational Statistics & Data Analysis, 53 (4), pp. 1400–1413.

GÓRECKI, T., KRZYŚKO, M., WASZAK, Ł., WOŁYŃSKI, W., (2014). Methods of reducing dimension for functional data, Statistics in Transition new series, 15, pp. 231–242.

GÓRECKI, T., KRZYŚKO, M., WOŁYŃSKI, W., (2016). Multivariate functional regression analysis with application to classification problems, In: Analysis of Large and Complex Data, Studies in Classification, Data Analysis, and Knowledge Organization, Eds.: Wilhelm Adalbert F. X., Kestler Hans A., Springer International Publishing, pp. 173–183.

GRETTON, A., BOUSQUET, O., SMOLA, A., SCHÖLKOPF, B., (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In: Algorithmic Learning Theory (S., Jain, H. U., Simon and E., Tomita, eds.), Lecture Notes in Computer Science, 3734, pp. 63–77, Springer, Berlin.

HASTIE, T. J., TIBSHIRANI, R. J., BUJA, A., (1995). Penalized discriminant analysis, Annals of Statistics, 23, pp. 73–102.

HORVÁTH, L., KOKOSZKA, P., (2012). Inference for Functional Data with Applications, Springer, New York.

JACQUES, J., PREDA, C., (2014). Model-based clustering for multivariate functional data, Computational Statistics & Data Analysis, 71, pp. 92–106.

KONG, J., WANG, S., WAHBA G., (2015). Using distance covariance for improved variable selection with application to learning genetic risk models, Statistics in Medicine, 34, pp. 1708–1720.

KUHN, M., Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt, (2018), caret: Classification and Regression Training. R package version 6.0-80, https://CRAN.R-project.org/package=caret.

R Core Team (2018). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

RAMSAY, J. O., SILVERMAN, B.W., (2005). Functional Data Analysis, Springer, New York.

RAMSAY, J. O., WICKHAM, H. GRAVES, S., HOOKER, G., (2018). fda: Functional Data Analysis, R package version 2.4.8, https://CRAN.R-project.org/package=fda.

RIZZO, M. L., SZÉKELY, G. J., (2018). energy: E-Statistics: Multivariate Inference via the Energy of Data, R package version 1.7-5, https://CRAN.R-project.org/package=energy.

ROSSI, F., DELANNAYC, N., CONAN-GUEZA, B., VERLEYSENC, M., (2005). Representation of functional data in neural networks, Neurocomputing, 64, pp. 183–210.

ROSSI, F., VILLA, N., (2006). Support vector machines for functional data classification, Neural Computing, 69, pp. 730–742.

ROSSI, N., WANG, X., RAMSAY, J.O., (2002). Nonparametric item response function estimates with EM algorithm, Journal of Educational and Behavioral Statistics, 27, pp. 291–317.

SCHÖLKOPF, B., SMOLA, A. J., MÜLLER, K. R., (1998). Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation, 10, pp. 1299–1319.

SZÉKELY, G. J., RIZZO, M. L., BAKIROV, N. K., (2007). Measuring and testing dependence by correlation of distances, The Annals of Statistics, 35 (6), pp. 2769–2794.

SZÉKELY, G. J., RIZZO, M. L., (2009). Brownian distance covariance, Annals of Applied Statistics, 3 (4), pp. 1236–1265.

SZÉKELY, G. J., RIZZO, M. L., (2012). On the uniqueness of distance covariance, Statistical Probability Letters, 82 (12), pp. 2278–2282.

SZÉKELY, G. J., RIZZO, M. L., (2013). The distance correlation t-test of independence in high dimension. Journal of Multivariate Analysis, 117, pp. 193–213.